DOCUMENT RESUME

ED 243 893 TM 830 259

AUTHOR Ward, William C.; And Others
TITLE Ill-Structured Problems as Multiple-Choice Items.
INSTITUTION Educational Testing Service, Princeton, N.J.; Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RR-83-6; GREB-81-18P
PUB DATE Mar 83
NOTE 28p.
AVAILABLE FROM Educational Testing Service, Research Publications R-116, Princeton, NJ 08541.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Higher Education; *Logical Thinking; *Multiple Choice Tests; Problem Solving; Test Construction; Testing Problems; *Test Items

ABSTRACT

A new item type was developed, incorporating features of "ill-structured" problems in a multiple-choice format. The problems are similar to previously developed scientific thinking tasks in requiring the examinee to go beyond the information provided; they resemble a variant of the logical reasoning item type, but demand somewhat more structuring by the examinee of the set of assumptions needed to solve the problem. A pretest was carried out to compare the new item type with two variants of the logical reasoning item type; a second study, using preexisting data, compared these two variants with one another. In neither case was there any indication that problems identified as "ill-structured" measured different aspects of ability than do "well-structured" problems. The new item type might be employed in tests of reasoning for the sake of increasing the variety of item types available for test construction, but would not be expected to extend the range of cognitive skills assessed by the tests. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED243893

TM 830 285

GRE

GRADUATE RECORD EXAMINATIONS

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. C. Ward, &

H. C. Weidenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ILL-STRUCTURED PROBLEMS AS MULTIPLE-CHOICE ITEMS

William C. Ward
Sybil B. Carlson
Erich Woisetschlaeger

GRE Board Professional Report GREB No. 81-18P

ETS Research Report 83-6

March 1983

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

GRE BOARD RESEARCH REPORTS FOR GENERAL AUDIENCE

Altman, R. A. and Wallmark, M. M. A Summary of Data from the Graduate Programs and Admissions Manual. 74-1R, 1975.

Baird, L. L. An Examination of the Graduate Study Application and Enrollment Decisions of GRE Candidates. 79-11R, 1982.

Baird, L. L. An Inventory of Documented Accomplishments. 77-3R, 1979.

Baird, L. L. Cooperative Student Survey (The Graduates [$2.50 each], and Careers and Curricula). 70-4R, 1973.

Baird, L. L. The Relationship Between Ratings of Graduate Departments and Faculty Publication Rates. 77-2aR, 1980.

Baird, L. L. and Knapp, J. E. The Inventory of Documented Accomplishments for Graduate Admissions: Results of a Field Trial Study of Its Reliability, Short-Term Correlates, and Evaluation. 78-3R, 1981.

Burns, R. L. Graduate Admissions and Fellowship Selection Policies and Procedures (Part I and II). 69-5R, 1970.

Centra, J. A. How Universities Evaluate Faculty Performance: A Survey of Department Heads. 75-5bR, 1977. ($1.50 each)

Centra, J. A. Women, Men and the Doctorate. 71-10R, 1974. ($3.50 each)

Clark, M. J. The Assessment of Quality in Ph.D. Programs: A Preliminary Report on Judgments by Graduate Deans. 72-7aR, 1974.

Clark, M. J. Program Review Practices of University Departments. 75-5aR, 1977. ($1.00 each)

Clark, M. J and Centra, J. A. Conditions Influencing the Career Accomplishments of Ph.Ds. 76-2R, 1982.

Donlon, T. F. Annotated Bibliography of Test Speededness. 76-9R, 1979.

Flaugher, R. L. The New Definitions of Test Fairness In Selection: Developments and Implications. 72-4R, 1974.

Fortna, R. O. Annotated Bibliography of the Graduate Record Examinations. 1979.

Frederiksen, N. and Ward, W. C. Measures for the Study of Creativity in Scientific Problem-Solving. 1978.

Hartnett, R. T. Sex Differences in the Environments of Graduate Students and Faculty. 77-2bR, 1981.

Hartnett, R. T. The Information Needs of Prospective Graduate Students. 77-8R 1979.

Hartnett, R. T. and Willingham, W. W. The Criterion Problem: What Measure of Success in Graduate Education? 77-4R, 1979.

Knapp, J. and Hamilton, I. B. The Effect of Nonstandard Undergraduate Assessment and Reporting Practices on the Graduate School Admissions Process. 76-14R, 1978.

Lannholm, G. V. and Parry, M. E. Programs for Disadvantaged Students in Graduate Schools. 69-1R, 1970.

Miller, R. and Wild, C. L. Restructuring the Graduate Record Examinations Aptitude Test. GRE Board Technical Report, 1979.

Powers, D. E. and Lehman, J. GRE Candidates' Perceptions of the Importance of Graduate Admission Factors. 81-2R, 1982.

Powers, D. E. and Swinton, S. S. Effects of Self-Study of Test Familiarization Materials for the Analytical Section of the GRE Aptitude Test. 79-9R, 1982.

Reilly, R. R. Critical Incidents of Graduate Student Performance. 70-5R, 1974.

Rock, D. and Werts, C. An Analysis of Time Related Score Increments and/or Decrements for GRE Repeaters across Ability and Sex Groups. 77-9R, 1979.

Rock, D. A. The Prediction of Doctorate Attainment in Psychology, Mathematics and Chemistry. 69-6aR, 1974.

Schrader, W. B. Admissions Test Scores as Predictors of Career Achievement in Psychology. 76-1aR, 1978.

Schrader, W. B. GRE Scores as Predictors of Career Achievement in History. 76-1bR, 1980.

Swinton, S. S. and Powers, D. E. A Study of the Effects of Special Preparation on GRE Analytical Scores and Item Types. 78-2R, 1982.

Wild, C. L., Swinton, S. S., and Wallmark, M. M. A Summary of the Research Leading to the Revision of the Format of the Graduate Record Examinations Aptitude Test in October 1981. 80-1aR, 1982.

Wild, C. L. Summary of Research on Restructuring the Graduate Record Examinations Aptitude Test. 1979.

Wild, C. L. and Durso, R. Effect of Increased Test-Taking Time on Test Scores by Ethnic Group, Age, and Sex. 76-6R, 1979.

Wilson, K. M. A Study of the Validity of the Restructured GRE Aptitude Test for Predicting First-Year Performance in Graduate Study. 78-6R, 1982.

Wilson, K. M. The GRE Cooperative Validity Studies Project. 75-8R, 1979.

Wiltsey, R. G. Doctoral Use of Foreign Languages: A Survey. 70-14R, 1972. (Highlights $1.00, Part I $2.00, Part II $1.50).

Witkin, H. A.; Moore, C. A.; Oltman, P. K.; Goodenough, D. R.; Friedman, F.; and Owen, D. R. A Longitudinal Study of the Role of Cognitive Styles in Academic Evolution During the College Years. 76-10R, 1977 ($5.00 each).

Ill-Structured Problems as Multiple-Choice Items

William C. Ward
Sybil B. Carlson
Erich Woisetschlaeger

GRE Board Professional Report GREB No. 81-18P

March 1983

Copyright © 1983 by Educational Testing Service. All rights reserved.

Abstract

A new item type was developed, incorporating features of "ill-structured" problems in a multiple-choice format. The problems are similar to previously developed scientific thinking tasks in requiring the examinee to go beyond the information provided; they resemble a variant of the logical reasoning item type, but demand somewhat more structuring by the examinee of the set of assumptions needed to solve the problem.

A pretest was carried out to compare the new item type with two variants of the logical reasoning item type; a second study, using preexisting data, compared these two variants with one another. In neither case was there any indication that problems identified as "ill-structured" measured different aspects of ability than do "well-structured" problems. The new item type might be employed in tests of reasoning for the sake of increasing the variety of item types available for test construction, but would not be expected to extend the range of cognitive skills assessed by the tests.

Ill-Structured Problems as Multiple-Choice Items

Purpose

The purpose of this research was to explore a new item type that incorporates features of "ill-structured" problems in a format suitable for inclusion in a standardized test. A "well-structured" problem is one whose solution requires only understanding and manipulating the information contained in the problem statement, often with the use of a familiar algorithm. Ill-structured problems (Simon, 1978), in contrast, are "those that (a) are more complex and have less definite criteria for determining when the problem has been solved, (b) do not provide all the information necessary to solve the problem, and (c) have 'no legal move generator' for finding all the possibilities at each step" (Frederiksen, in process). Calculating the volume of a room, given length, breadth, and height, would constitute a well-structured problem; determining the most functional arrangement of the furniture within it would be an ill-structured one.

This work grew out of research on scientific problem solving that was conducted with the support of the Graduate Record Examinations Board (Frederiksen & Ward, 1978; Ward, Frederiksen, & Carlson, 1980). In that research four tests were developed, each measuring a kind of problem solving that might be required of a behavioral scientist in conducting and interpreting research studies. In the Formulating Hypotheses Test, for example, the examinee was given a brief description of a research investigation, a table or graph presenting the results of the study, and a one-sentence statement of the major finding. The task was to write the hypothesis most likely to provide the correct explanation of the finding, along with competing hypotheses that should be considered in interpreting the results or in planning further research.

The scientific thinking items are examples of ill-structured problems. They draw on a set of abilities related, but not identical, to those that determine performance on standardized aptitude and achievement tests; the true score correlation of Formulating Hypotheses with the GRE Advanced Psychology Test, for example, is about .5. Thus, there is reason to think that inclusion of measures of this sort in standardized examinations could increase the breadth of measurement obtained and, perhaps, could improve the validity of the examinations by providing problems representative of situations faced in scientific and professional activities.

However, such free-response measures could not easily be incorporated into large-scale testing, at least within current limitations on the time available for testing and for scoring. An effort was made in earlier work (Ward et al., 1980) to

create machine-scorable versions of these problems; but the format chosen proved to require substantial time per item, to produce relatively unreliable scores, and to be susceptible to possible response-bias effects on the scores. The present investigation is an attempt to overcome these limitations and to produce ill-structured problems that would be suitable for inclusion in a standardized aptitude test.

Description of Problems

Appendix A provides examples of the items developed in this study. These problems differ in several ways from the earlier scientific thinking problems. First, they require no special technical knowledge. While the examinee must go beyond the information given, he or she does so primarily in order to determine what assumptions are required to link an argument to a conclusion or what inferences might plausibly be drawn from a finding. Second, they employ a standard five-option multiple-choice format. Third, they require much less time per item than the scientific thinking problems. An item can be completed in about 1.5 minutes, permitting a larger number and greater diversity of items to be presented in a given time period.

In other respects, the problems are very similar to those employed earlier. Each problem statement consists of a graph or table sometimes accompanied by three to five lines of text outlining a situation, along with a brief verbal statement of a finding--a conclusion that can be made from the information displayed. The examinee is then asked to answer a question dealing with possible interpretations or explanations of the finding. For example:

> Which of the following, if true, is most likely to explain the finding above?
>
> One possible explanation of the finding is A. To evaluate this explanation, it would be useful to know each of the following EXCEPT:
>
> One possible interpretation of the finding is X. One could argue against this interpretation most effectively by pointing out:

Project Activities

Four activities were completed in the course of this investigation. First, a trial set of items was written and revised. Second, a comparison was made of several existing

variants of the logical reasoning item type. Third, a pretest was carried out, including a comparison of the new items with these logical reasoning item types. Finally, the new items were reviewed by test development experts.

Item writing and review. A small preliminary set of items was written, both to sharpen the initial conception of the item type and to provide material for use in talk-aloud sessions. Several individuals (undergraduate students and college graduates) verbalized their reasoning as they attempted to solve the problems; their statements served to identify ambiguities and to suggest the kinds of inferences and assumptions it is feasible to require of examinees. A member of the ETS test development staff with extensive experience in writing reasoning items critiqued the draft items and provided general suggestions on the art of item writing. A set of 10 items was prepared, critiqued, revised, and judged to provide a good representation of the item type.

Eight test development experts reviewed the items and indicated which option they would choose as correct; thirty-seven undergraduate students attempted the items. Information obtained from both groups served as the basis for further revisions.

Examination of logical reasoning items. Once a few of the new items had been written, it became clear that they had a strong resemblance to a variant of the logical reasoning item type. Logical reasoning items that have appeared in disclosed GRE tests have tended to be very tightly constrained deductive problems--and, indeed, could serve as prototypes of well-structured problems (see Appendix B). In contrast, this variant includes items that require substantial structuring of the problem by the examinee (see Appendix C). Following a statement of facts or an argument, the examinee is asked to respond to questions like:

> Which, of the following, if true, would most weaken this claim?
>
> The author's argument would be strengthened if it were discovered that:

These items overlap with the new items we have developed in requiring the examinee to produce his or her own assumptions or interpretations in order to answer the question and to compare all the alternatives to determine which is the most relevant to the situation described. There are differences in the degree to which the examinee must provide structure to solve the problem, as well as differences in format, but the two item types seem far more similar to one another than either is to the typical well-structured logical reasoning item.

One consequence of this convergence is that, if the new ill-structured problems should prove to have desirable measurement characteristics, it is clear how the item type could be incorporated into the GRE tests. Its use would require only a revision of specifications for the present analytical section of the General Test, so as to offer a greater variety of items within the general category of logical reasoning.

A second consequence is that there seemed little purpose in writing a larger set of the new problems. If more are desired for further investigation, it should be feasible and efficient to ask ETS test development staff to produce them.

Finally, past experience with what we will refer to as ill-structured logical reasoning problems should be relevant to assessing the possibility of writing examples of the new items that will provide good measurement.

Prior to its 1981 revision, the GRE General Test contained only four logical reasoning items in each test form. Disclosed GRE tests, therefore, contain too few items to permit a comparison of variants of the item type. The Law School Admission Test, however, can provide relevant information, since that test has for some time contained a 25-item logical reasoning section.[1]

Logical reasoning items from six disclosed forms of the LSAT were classified into two categories. "Ill-structured" items were those 18 that two judges agreed were clearly examples of this variant; "other" items were the 100 that were either definitely well-structured or ambiguous in their classification. Within each form, for each category, the median item correlation with the total logical reasoning score for the test was obtained. When the medians of these medians were obtained, it was evident that the two sets of items showed little difference in their relation to the criterion--the median was .48 for those classified as ill-structured, and .46 for all others.

Two conclusions can be drawn. First, when judged against a narrowly relevant criterion, ill-structured logical reasoning items appear to function well. Second, however, there is some reason to doubt whether they yield very distinctive measurement from that provided by well-structured logical reasoning items. The criterion against which they were judged is a score predominantly derived from performance on well-structured items, and yet the ill-structured items correlate as well with that criterion as do the well-structured ones.

---

[1]Thanks are due to Thomas O. White and Franklin R. Evans of the Law School Admission Services for providing access to LSAT items and item statistics.

Pretest. The results of the analysis of LSAT items suggested the desirability of a pretest that not only would examine characteristics of the new items in isolation, but would also allow more thorough study of their relations to existing item types. A pretest was conducted with 68 undergraduates, who completed four short tests consisting of (a) the set of 10 new ill-structured problems; (b) 10 ill-structured logical reasoning items, revised from items developed and pretested for possible use in the GRE analytical section; (c) 10 well-structured logical reasoning items, drawn from items employed in disclosed forms of the GRE General Test; and (d) 20 letter sets items. Letter sets were included to allow a comparison of the relation of the first three item types mentioned to a score representing inductive reasoning, a factor correlated with but distinguishable from logical reasoning.

Descriptive results of standard item analyses are shown in Table 1. The 10-item test comprised of new ill-structured items was appropriate in difficulty for the sample, with a mean number right score of 4.75, and with the percent correct on individual items ranging from 31 to 76. Biserial correlations with the total score ranged from .40 to .83, and the reliability of the test was .62 (coefficient alpha). Results for the remaining two logical reasoning item types were quite similar to these, while the longer letter sets test was somewhat easier and more reliable.

Of principal interest are the correlations among tests employing the various item types, shown in Table 2. Zero-order correlations are presented above the main diagonal, while correlations corrected for the unreliability of the measures are below the diagonal. Three aspects of these relationships are worthy of note. First, the correlations among the first three tests are very similar to one another. In terms of uncorrected coefficients, a test consisting of the new ill-structured problems correlates .73 and .68, respectively, with one comprised of GRE ill-structured items and one made up of well-structured items; these latter two correlate .69 with one another. Second, when corrected for unreliability, these coefficients suggest that the three item types overlap essentially completely in what they measure. The corrected coefficients, in fact, all exceed 1.0--an "impossible" result that arises because the reliability estimate used in making the correction is a lower-bound estimate, producing conservative values for reliabilities but, when employed as a divisor in the correction formula, overestimates of true relationships. Third, relationships to the letter sets test are nearly identical for the three logical reasoning tests. The uncorrected coefficients range only from .57 to .58, while the corrected ones vary from .74 to .80. Logical reasoning tests in general can be distinguished

Table 1

Pretest Results: Descriptive Statistics

Item Type	n of Items	Mean	Standard Deviation	Median r Biserial	Reliability
New Ill-Structured	10	4.75	2.26	.62	.62
GRE Ill-Structured	10	5.26	2.10	.57	.57
GRE Well-Structured	10	6.10	2.34	.73	.68
Letter Sets	20	14.24	4.92	.82	.89

Table 2

Pretest Results: Correlations Among Tests[1]

Item Type		(1)	(2)	(3)	(4)
New Ill-Structured	(1)		.73	.68	.58
GRE Ill-Structured	(2)	1.23		.69	.57
GRE Well-Structured	(3)	1.06	1.10		.58
Letter Sets	(4)	.78	.80	.74	

[1]Zero-order coefficients above main diagonal, correlations corrected for unreliability below. N = 68.

from the inductive reasoning represented by the letter sets test, but there is no basis in these data for suggesting that the new ill-structured problems measure anything different from the abilities drawn upon by logical reasoning item types that have previously been employed.

Review. Five test development experts reviewed the final versions of the new ill-structured items; none of these individuals had been involved in the earlier reviews. Their responses and comments suggested that about five of the ten items could meet standards of defensibility and lack of ambiguity appropriate for operational use (although, even for these, minor revisions might be needed), three would require extensive reworking, and the final two would probably be discarded. It was both disappointing and surprising that the review revealed so many difficulties: disappointing because of the thoroughness of the critique and revision that had preceded this examination of the items, surprising in view of the very satisfactory item statistics obtained in the pretest described above.

Interpretation of this result is difficult. To some extent it may reflect problems inherent in learning to write a somewhat different item type than has previously been used, rather than any ambiguities intrinsic to the format. This suggestion is strengthened by the similarity between the new items and items that have passed successfully through standard review procedures and have been included in operational examinations. On the other hand, it may not be unreasonable to expect somewhat greater difficulties in achieving unequivocally defensible keys for new types of items, to the degree that those items require the use of general knowledge and the development of assumptions that are not strictly entailed by the item stem. Perhaps an appropriate conclusion is that the question will not be answered in a research investigation; only experience obtained as test development staff attempt to produce items for operational use will indicate the magnitude of the difficulty in writing acceptable items.

Conclusions

It is feasible to write multiple-choice reasoning items that embody some of the features of ill-structured problems. Some existing variants of logical reasoning problems represent a step in this direction, while the problems developed for the present investigation represent a further step.

From the present research, however, there is no evidence that such items provide measurement of abilities different from those underlying performance on well-structured reasoning problems. Rather,

logical reasoning items ranging from well- to ill-structured appear to measure substantially the same set of abilities. Items like those developed for this study might be of interest as a way of broadening the set of item formats on which test developers can draw in constructing tests of reasoning, but apparently not as a way of extending the range of cognitive skills assessed by the tests.

From our own empirical work and from theoretical writings such as those of Simon (1978), we are convinced that there are sometimes differences between well- and ill-structured problems. Why, then, did differences not appear here? A possible explanation may lie in the role of knowledge in problem solving. Real-world ill-structured problems often cannot be solved through the use of general knowledge and skills alone, but rather require specific information and experience that may be available to one individual but not to another. In the example with which this report began, the ill-structured problem of how to arrange the furniture within a room might be solved very differently by the interior decorator who has a broad store of relevant knowledge on which to draw and by the tyro moving into his or her first apartment. In the context of aptitude testing, posing this question to both the expert and the novice would be seen as unfair; yet recasting the problem so as to eliminate any effect of special experience might change it importantly. Perhaps a similar effect is occurring in the present investigation--in the effort to avoid requiring knowledge that would be differentially available to different individuals, an important part of what makes a problem ill-structured is lost, and more general aptitudes play the major part in determining task performance.

This reasoning suggests that further investigation of ill-structured problems should be carried out in a context of achievement rather than aptitude testing. Only in such a setting will we be willing to employ items that require specific knowledge that the examinee must possess, identify as relevant, and use appropriately in problem solution.

From a longer-term perspective, it may also be worthwhile to consider approaches to the assessment of problem solving that yield for each item more than a single score of "right" or "wrong." Rather than constraining the problem to meet the limitations of a multiple-choice format, it may be desirable to work toward methods of presentation that allow multiple responses by an examinee and methods of scoring that distinguish among the skills, knowledge, and abilities that contribute to problem solution. An illustration is provided by research in which an attempt was made to model the process of medical diagnostic problem solving; each problem consisted of several cycles of

information gathering and hypothesis generation, permitting the derivation of scores reflecting multiple aspects of performance (Frederiksen, Ward, Case, Carlson, & Samph, 1981). If such work could be combined with insights from cognitive science research on the nature of the cognitive "components" and "metacomponents" underlying intellectual performance (e.g., Sternberg, 1981), more informative and face-valid assessments of problem solving might be achieved.

References

Frederiksen, N. Implications of theory for instruction in problem solving. ETS Research Report. In process.

Frederiksen, N., & Ward, W. C. Measures for the study of creativity in scientific problem solving. Applied Psychological Measurement, 1978, 2, 1-24.

Frederiksen, N., Ward, W. C., Case, S. M., Carlson, S. B., & Samph, T. Development of methods for selection and evaluation in undergraduate medical education (ETS RR 81-4). Princeton, N.J.: Educational Testing Service, 1981.

Simon, H. A. Information-processing theory of human problem solving. In W. K. Estes (Ed.), Handbook of learning and cognitive processes (Vol. 5), Human information processing. Hillsdale, N.J.: Erlbaum, 1978.

Sternberg, R. J. Testing and cognitive psychology. American Psychologist, 1981, 36, 1181-1189.

Ward, W. C., Frederiksen, N., & Carlson, S. B. Construct validity of free-response and machine-scorable forms of a test. Journal of Educational Measurement, 1980, 17, 11-29.

Appendix A

Examples of New Ill-Structured Problems

Year-end Grades in Ms. Francoise's French Class:

Number of Students Getting Each Grade

	Girls	Boys
A	6	0
B	5	6
C	7	9
D	0	4
F	0	3

Finding: On the average, girls in Ms. Francoise's class obtained much higher grades than did boys.

John, whose grade was a D, complained that the teacher was biased in giving higher grades to girls. Alice, who got an A, objected. Alice could argue against John most effectively by pointing out which of the following:

*(A) Every girl in the class spent at least three hours each week doing French homework, while most of the boys spent one hour or less each week.

(B) The difference in grades for boys and girls this year is similar to differences found in each of the eight years Ms. Francoise has taught French.

(C) To determine a student's grade, Ms. Francoise gave more weight to class participation and weekly oral quizzes than to written examinations.

(D) Not even one of the girls obtained a grade lower than C, while none of the boys obtained an A.

(E) In Mr. Chofsky's Russian class, boys received higher average grades than did girls in each of the last five years.

Schizophrenia in Identical Twins

A study was carried out investigating the mental health of 2000 pairs of identical twins, all of whom were reared by their natural parents. The incidence of schizophrenia in this sample was as follows:

Mental Health Status	Number of Cases
Both Twins Schizophrenic	87
One Twin Schizophrenic	8
Neither Twin Schizophrenic	1905

Finding: If one member of a pair of identical twins was schizophrenic, there was a very high probability that the other was also.

One explanation of this finding is that schizophrenia is a genetically determined disease. To evaluate this explanation, it would be useful to know each of the following EXCEPT:

(A) incidence of schizophrenia in identical twins separated at birth and reared in different homes

(B) incidence of schizophrenia in twins that are not identical

(C) incidence of schizophrenia in unrelated children reared in the same household

*(D) incidence of schizophrenia in children under the age of 6

(E) incidence of schizophrenia in groups in which there are a significant number of intermarriages

Household Incomes for Different Age Groups

A recent government publication cited U.S. population figures by age and average household income. Household incomes were calculated as a percent of the national average. Income includes money available from all sources--wages and salaries, investments, pensions, Social Security, and so forth.

Age of Head of Household	Household Income as Percent of National Average
65+	69%
55-64	120%
45-54	135%
35-44	126%
25-34	92%
15-24	85%

Finding: The average income is lower for households headed by people over 65 than for households headed by people in other age groups.

A possible conclusion is that those living in households headed by people over 65 are more likely than those in other groups to live under substandard conditions. To evaluate this conclusion, which of the following additional items of information would be the most useful to have?

(A) In each age group, the percent of household income that is being saved

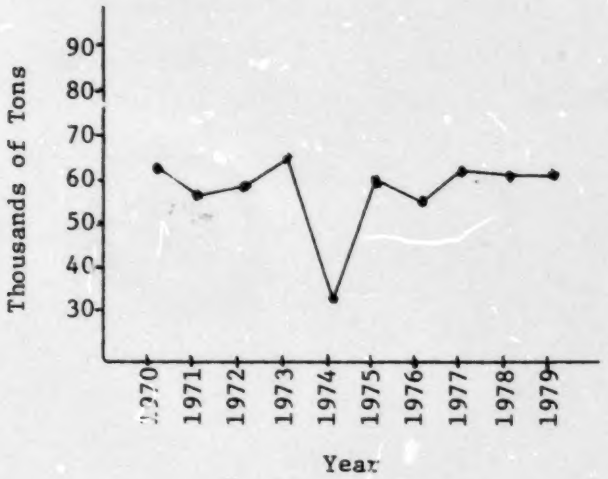
*(B) In each age group, the average number of people per household

(C) The levels of income people over age 65 had maintained before reaching 65

(D) The percent of people over 65 who live in households headed by people over 65

(E) Average incomes for households in which the head of household is over 65 and continues to have full-time employment

Annual Mackerel Catch by Fleet Sailing from Port Byardia



Finding: The Port Byardia fleet had a mackerel catch that was relatively constant year-to-year during the 1970's, except for a sharp drop in 1974.

Each of the following, if true, is a plausible explanation for the poor catch in 1974 EXCEPT:

(A) A major oil spill during the 1974 fishing season temporarily depleted the mackerel's food supply, leading them to change their feeding grounds for the remainder of the year.

(B) A prolonged strike at the only cannery near Port Byardia eliminated the fishermen's outlet for sale of their catch, so they stopped fishing halfway through the 1974 season.

(C) During 1974, boats of other nations competed on grounds that had previously been fished only by those from Port Byardia; by the next season a treaty had been negotiated reserving those grounds for the local fishermen.

*(D) Outmoded equipment and methods cut down on the fleet's effectiveness; after the 1974 season, more modern equipment was introduced and allowed a return to previous levels of success.

(E) Unusually severe storms cut drastically the number of days boats were able to fish during 1974.

Effectiveness of a New Drug

A doctor specializing in the treatment of insomnia decided to test the effects of a new antidepressant drug on his patients. He gave the drug to 20 of his patients, while to another 20 he gave placebos (sugar pills). One week after giving the pills to each patient, he asked what the effect had been on their sleeping. The patients' reports were as follows:

	Dramatic Improvement	Some Improvement	No Change
Patients Given New Drug	14	4	2
Patients Given Placebo	4	3	13

Finding: Almost all of the patients given the new drug reported improvement in their sleeping, while the majority of those who received placebos reported no improvement.

Before you could conclude that the new drug is effective in treating insomnia, you would need to be sure that the results were not due to a flaw in the way the test was carried out. Each of the following, if true, would provide a strong reason to doubt the effectiveness of the new drug EXCEPT:

(A) The doctor showed his enthusiasm for the new drug when he prescribed it to patients; their expectations, rather than the drug itself, determined how well they slept for the next week.

*(B) The doctor told all 40 patients that they were receiving an experimental drug; those who didn't like the idea became anxious and failed to show improvement in their sleeping.

(C) The doctor was biased in assigning drugs and placebos to patients, giving the new drug to patients with mild cases of insomnia and the placebo to those with more severe insomnia.

(D) The patients were aware of whether they were receiving the drug or a placebo, and reported to the doctor the results they thought he wanted to hear.

(E) The doctor was biased in recording the patients' reports, exaggerating positive statements made by patients receiving the new drug and minimizing positive statements made by those receiving placebos.

Appendix B

Examples of Well-Structured Logical Reasoning Problems

To be a good debater, one must be intelligent. Some good debaters, however, are also contentious, and contentious persons are always boring.

Which of the following conclusions can be properly drawn from the statements above?

(A) All good debaters are boring.

(B) All contentious persons are good debaters.

(C) Only good debaters are contentious

*(D) Some intelligent persons are boring.

(E) Most intelligent persons are boring.

Most people who stop smoking gain weight; therefore, if Mark does not stop smoking, he will probably not gain weight.

The argument above is most like which of the following?

(A) Most spiders live in dry places; therefore, spiders are probably found in all deserts.

(B) Most crimes of violence have increased in number during the past few years; therefore, law enforcement during this period has become lax.

(C) Most new sports cars are equipped with radial tires; therefore, Eve's new sports car may not be equipped with radial tires.

(D) Most countries ruled by dictators do not have a free press; therefore, the nation of Nogub, which is ruled by a dictator, probably does not have a free press.

*(E) Most popular television programs are of low intellectual quality; therefore, "Fun Fair," an unpopular program, is probably of high intellectual quality.

There is no reason to rule out the possibility of life on Uranus. We must, then, undertake the exploration of that planet.

The argument above assumes that

(A) life exists on Uranus

(B) Uranus is the only other planet in the solar system capable of supporting life

(C) Uranian life would be readily recognizable as life

*(D) the search for life is a sufficient motive for space exploration

(E) no one has previously proposed the exploration of Uranus

If Ramón was born in New York State, then he is a citizen of the United States.

The statement above can be deduced logically from which of the following statements?

*(A) Everyone born in New York State is a citizen of the United States.

(B) Every citizen of the United States is a resident either of one of the states or of one of the territories.

(C) Some people born in New York State are citizens of the United States.

(D) Ramón was born either in New York or in Florida.

(E) Ramón is a citizen either of the United States or of the Dominican Republic.

Appendix C

Examples of Ill-Structured Logical Reasoning Problems

From studies of 108 South American mummies, archeologists have concluded that pneumonia was a major cause of death in that continent 3,000 years ago, and that the incidence of death due to pneumonia was close to the rate of death by pneumonia today.

Which of the following would be most important in evaluating the accuracy of the conclusion above?

(A) The incidence of pneumonia death as evidenced by mummies from other continents

(B) Whether people 3,000 years ago had any conception of the disease pneumonia

(C) Causes of death, other than pneumonia, apparent in the mummies studied

(D) The general incidence of disease in South America 3,000 years ago

*(E) If the mummies were representative of the South American population of their time

An accident that occurred recently at a nuclear reactor has served to remind us that, ever since the discovery of fire, people have occasionally burned their hands. Some naysayers may now demand an end to the production of nuclear fuel, but that is the wrong approach. We must not return to the Stone Age. The price of civilization is that we must risk a few burnt hands if the fires are to be kept burning.

Which of the following, if true, most weakens the argument above?

(A) Whereas fire is accessible to all nations and virtually all individuals, nuclear fuel is accessible only to limited numbers of nations and individuals.

(B) The production of nuclear fuel is more expensive than the production of fire.

(C) The process by which nuclear fuel is produced is less easily accomplished than is the process by which fire is produced.

(D) Fire has a wider range of uses than does nuclear fuel.

*(E) The risks involved in the production of nuclear fuel are much more serious than those involved in the production of fire.

In the United States, rumors are springing up at a faster rate than ever before, and at the same time people are becoming more willing to accept rumors. The reason is not hard to uncover: Americans just do not know what to believe these days or whom, since their leaders command less credibility than they once did.

Each of the following could logically be a factor contributing to the situation described in the passage above EXCEPT:

(A) People have a psychological need to have reasonably definite beliefs about areas of public life that they care about.

*(B) Contemporary American thinking is predominantly skeptical, doubting everything and believing nothing.

(C) The more favorable conditions are for rumors to gain currency, the more rumors will be put into circulation.

(D) Isolated rumors are relatively easy to resist; a multitude of rumors compels acceptance of at least some, on the theory that where there is smoke, there must be fire.

(E) Modern civilization is so complex that political leaders often have a less than adequate understanding of the social, economic, and political questions that demand policy decisions.

Cities should pass ordinances that require municipal employees to live in the city for which they work. Such requirements would give the employees a better understanding of the needs of the community they serve and hence would improve services. The employees would also be nearer their jobs, and for public servants like police officers and firefighters, this proximity is crucial.

Which of the following, if true, is the best criticism of this proposal to improve city services?

(A) Cities may want to impose residency requirements retroactively on persons who already serve the city but live elsewhere.

(B) Cities may differ according to how they define residency or how much time they require to establish residency.

*(C) Residency requirements will substantially shrink the pool of qualified persons from which the city can recruit employees.

(D) A municipal employee may have a job that takes him or her to many different areas of the city.

(E) Residents of small cities would be close to their job sites, but residents of large cities could still live far from theirs.

GRE BOARD RESEARCH REPORTS OF A TECHNICAL NATURE

Boldt, R. R. Comparison of a Bayesian and a Least Squares Method of Educational Prediction. 70-3P, 1975.

Campbell, J. T. and Belcher, L. H. Word Associations of Students at Predominantly White and Predominantly Black Colleges. 71-6P, 1975.

Campbell, J. T. and Donlon, T. F. Relationship of the Figure Location Test to Choice of Graduate Major. 75-7P, 1980.

Carlson, A. B.; Reilly, R. R.; Mahoney, M. H.; and Casserly, P. L. The Development and Pilot Testing of Criterion Rating Scales. 73-1P, 1976.

Carlson, A. B.; Evans, F.R.; and Kuykendall, N. M. The Feasibility of Common Criterion Validity Studies of the GRE. 71-1P, 1974.

Centra, J. A. Graduate Degree Aspirations of Ethnic Student Groups Among GRE Test-Takers. 77-7P, 1980.

Donlon, T. F. An Exploratory Study of the Implications of Test Speededness. 76-9P, 1980.

Donlon, T. F.; Reilly, R. R.; and McKee, J. D. Development of a Test of Global vs. Articulated Thinking: The Figure Location Test. 74-9P, 1978.

Echternacht, G. Alternate Methods of Equating GRE Advanced Tests. 69-2P, 1974.

Echternacht, G. A Comparison of Various Item Option Weighting Schemes/A Note on the Variances of Empirically Derived Option Scoring Weights. 71-17P, 1975.

Echternacht, G. A Quick Method for Determining Test Bias. 70-8P, 1974.

Evans, F. R. The GRE-Q Coaching/Instruction Study. 71-5aP, 1977.

Fredericksen, N. and Ward, W. C. Development of Measures for the Study of Creativity. 72-2P, 1975.

Kingston, N. and Dorans, N. Effect of the Position of an Item Within a Test on Item Responding Behavior: An Analysis Based on Item Response Theory. 79-12bP, 1982.

Kingston, N. M. and Dorans, N. J. The Feasibility of Using Item Response Theory as a Psychometric Model for the GRE Aptitude Test. 79-12P, 1982.

Levine, M. V. and Drasgow, F. Appropriateness Measurement with Aptitude Test Data and Esimated Parameters. 75-3P, 1980.

McPeek, M.; Altman, R. A.; Wallmark, M.; and Wingersky, B. C. An Investigation of the Feasibility of Obtaining Additional Subscores on the GRE Advanced Psychology Test. 74-4P, 1976.

Oltman, P. K. Content Representativeness of the Graduate Record Examinations Advanced Tests in Chemistry, Computer Science, and Education. 81-12P, 1982.

Pike, L. Implicit Guessing Strategies of GRE Aptitude Examinees Classified by Ethnic Group and Sex. 75-10P, 1980.

Powers, D. E.; Swinton, S.; Thayer, D.; and Yates, A. A Factor Analytic Investigation of Seven Experimental Analytical Item Types. 77-1P, 1978.

Powers, D. E.; Swinton, S. S.; and Carlson, A. B. A Factor Analytic Study of the GRE Aptitude Test. 75-11P, 1977.

Reilly, R. R. and Jackson, R. Effects of Empirical Option Weighting on Reliability and Validity of the GRE. 71-9P, 1974.

Reilly, R. R. Factors in Graduate Student Performance. 71-2P, 1974.

Rock, D. A. The Identification of Population Moderators and Their Effect on the Prediction of Doctorate Attainment. 69-6bP, 1975.

Rock, D. A. The "Test Chooser": A Different Approach to a Prediction Weighting Scheme. 70-2P, 1974.

Rock, D., Werts, C., and Grandy, J. Construct Validity of the GRE Aptitude Test Across Populations--An Empirical Confirmatory Study. 78-1P, 1982.

Sharon, A. T. Test of English as a Foreign Language as a Moderator of Graduate Record Examinations Scores in the Prediction of Foreign Students' Grades in Graduate School. 70-1P, 1974.

Stricker, L. J. A New Index of Differential Subgroup Performance: Application to the GRE Aptitude Test. 78-7P, 1981.

Swinton, S. S. and Powers, D. E. A Factor Analytic Study of the Restructured GRE Aptitude Test. 77-6P, 1980.

Ward, W. C. A Comparison of Free-Response and Multiple-Choice Forms of Verbal Aptitude Tests. 79-8P, 1982.

Ward, W. C.; Frederiksen, N.; and Carlson, S. B. Construct Validity of Free-Response and Machine-Scorable Versions of a Test of Scientific Thinking. 74-8P, 1978.

Ward, W. C. and Frederiksen, N. A Study of the Predictive Validity of the Tests of Scientific Thinking. 74-6P, 1977.

Wild, C. L., Swinton, S. S., and Wallmark, M. M. Research Leading to the Revision of the Format of the Graduate Record Examinations Aptitude Test in October 1981. 80-1bP, 1982.

END

DEPT. OF EDUCATION

NAT'L INSTITUTE OF EDUCATION

ERIC

DATE FILMED

SEPTEMBER

7 _1984